数据科学系列课程

——数据挖掘实战课程



巴特尔·马哈比尔 工程师|架构师|数据科学家|技术写作者

Bater.Makhabel (<http://linkedin.com/in/batermj>)

电话：13910817749，微信号：DR\_BATER\_MAKHABEL

编辑历史：

|  |  |  |
| --- | --- | --- |
| 版本 | 内容 | 日期 |
| V1.0 | 数据科学系列课程之数据挖掘实战 | 2018年2月2日 |
|  |  |  |

内容  
大纲

1. 数据挖掘引论 1

数据挖掘的由来 1

数据挖掘的现状 1

数据挖掘的分类 1

数据挖掘的将来 1

开发环境的配置 1

分布式运行环境的配置 1

docker和k8s 1

Python和数据挖掘 1

R和数据挖掘 1

分布式环境和数据挖掘 1

实战案例 1

总结 1

习题（在线题库） 1

2. 数据挖掘工程 1

数据挖掘工程方法论 2

数据挖掘工程流程 2

实战案例 2

总结 2

习题（在线题库） 2

3. 数据预处理算法 2

理论 2

实战案例 2

总结 2

习题（在线题库） 2

4. 数据可视化 2

理论 3

实战案例 3

总结 3

习题（在线题库） 3

5. 特征工程算法 3

特征工程介绍 3

特征的选择 3

特征的抽取 3

特征选择评价函数 3

线性特征变换 3

实战案例 3

总结 3

习题（在线题库） 3

6. 数据挖掘算法 3

介绍 4

算法与数据挖掘 4

数据挖掘算法的分类 4

数据挖掘算法的评估 4

数据挖掘算法的大规模分布式实现 4

频繁模式挖掘Frequent Pattern Mining 4

FP-Growth 4

Constraint-Based Frequent Pattern Mining 4

Mining High-Dimensional Data and Colossal Patterns 4

Mining Compressed or Approximate Patterns 5

Complete Pattern 5

Close Pattern 5

PrefixSpan 5

分类算法 6

决策树 6

基于规则的决策树 6

近邻分类算法 6

Logistic Regression 7

最大熵模型分类算法 7

贝叶斯分类器 7

Bayesian Belief Networks 8

EM 8

聚类算法 8

Partitioning Methods 8

Hierarchical Methods 9

Density-Based Methods 9

Grid-Based Methods 9

Probabilistic Model-Based Clustering 10

Clustering High-Dimensional Data 10

Clustering Graph and Network Data 10

Clustering with Constraints 11

KMeans算法 11

高斯混合 11

快速迭代聚类 12

LDA 12

Outlier Detection 12

Statistical Approaches 12

Proximity-Based Approaches 13

Clustering-Based Approaches 13

Classification-Based Approaches 13

Mining Contextual and Collective Outliers 14

Outlier Detection in High-Dimensional 14

实战案例 15

Spam Filtering 15

基于回归预测PV 15

社交网络图分析 15

信用卡欺诈检测 15

恶意软件智能检测 15

安全审计踪迹分析 15

入侵检测分析 15

计算机取证系统 15

口碑技术Online Reputation 15

总结 15

习题（在线题库） 15

7. 图数据挖掘算法 15

理论 16

实战案例 16

总结 16

习题（在线题库） 16

8. 社交网络数据挖掘算法 16

理论 16

实战案例 16

总结 16

习题（在线题库） 16

9. 多媒体数据挖掘算法 17

理论 17

实战案例 17

总结 17

习题（在线题库） 17

10. 文本数据挖掘算法 17

理论 18

话题挖掘 18

多文档自动文摘 18

情感分析 18

新闻摘要 18

实战案例 18

总结 18

习题（在线题库） 18

11. Web数据挖掘算法 18

理论 18

实战案例 18

总结 18

习题（在线题库） 18

作者简介 20

个人简介 21

作品封面 21

课程报名方式 25

开课时间 25

授课方式 25

收费方式和时间 25

报名方式 26

报名方式之支付宝 26

报名方式之LinkedIn 28

报名方式之微博 30

报名方式之邮件 32

报名方式之微信 32

报名方式之微信公众号 34

# 数据挖掘引论

## 数据挖掘的由来

## 数据挖掘的现状

## 数据挖掘的分类

## 数据挖掘的将来

## 开发环境的配置

## 分布式运行环境的配置

## docker和k8s

## Python和数据挖掘

## R和数据挖掘

## 分布式环境和数据挖掘

## 实战案例

## 总结

## 习题（在线题库）

# 数据挖掘工程

## 数据挖掘工程方法论

## 数据挖掘工程流程

## 实战案例

## 总结

## 习题（在线题库）

# 数据预处理

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 数据探索

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 数据可视化

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 特征工程算法

## 特征工程介绍

## 特征的选择

## 特征的抽取

## 特征选择评价函数

## 线性特征变换

## 实战案例

## 总结

## 习题（在线题库）

# 数据挖掘算法

## 介绍

### 算法与数据挖掘

### 数据挖掘算法的分类

### 数据挖掘算法的评估

### 数据挖掘算法的大规模分布式实现

## 频繁模式挖掘Frequent Pattern Mining

### FP-Growth

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Constraint-Based Frequent Pattern Mining

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Mining High-Dimensional Data and Colossal Patterns

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Mining Compressed or Approximate Patterns

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Complete Pattern

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Close Pattern

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### PrefixSpan

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

## 分类算法

### 决策树

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 基于规则的决策树

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 近邻分类算法

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Logistic Regression

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 最大熵模型分类算法

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 贝叶斯分类器

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Bayesian Belief Networks

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### EM

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

## 聚类算法

### Partitioning Methods

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Hierarchical Methods

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Density-Based Methods

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Grid-Based Methods

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Probabilistic Model-Based Clustering

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Clustering High-Dimensional Data

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Clustering Graph and Network Data

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Clustering with Constraints

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### KMeans算法

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 高斯混合

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### 快速迭代聚类

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### LDA

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

## Outlier Detection

### Statistical Approaches

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Proximity-Based Approaches

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Clustering-Based Approaches

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Classification-Based Approaches

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Mining Contextual and Collective Outliers

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

### Outlier Detection in High-Dimensional

#### 理论基础

#### 思路扩展

#### 并行算法

#### 分布式算法

## 实战案例

### Spam Filtering

### 基于回归预测PV

### 社交网络图分析

### 信用卡欺诈检测

### 恶意软件智能检测

### 安全审计踪迹分析

### 入侵检测分析

### 计算机取证系统

### 口碑技术Online Reputation

## 总结

## 习题（在线题库）

# 图数据挖掘算法

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 社交网络数据挖掘算法

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 多媒体数据挖掘算法

## 理论

## 实战案例

## 总结

## 习题（在线题库）

# 文本数据挖掘算法

## 理论

### 话题挖掘

### 多文档自动文摘

### 情感分析

### 新闻摘要

## 实战案例

## 总结

## 习题（在线题库）

# Web数据挖掘算法

## 理论

## 实战案例

## 总结

## 习题（在线题库）